Book Reviews

Advances in Large Margin Classifiers—A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans (MIT Press, 2000, series on Neural Information Processing, 422 pp., hardbound, ISBN 0-262-19 448-1) *Reviewed by Irwin King.*

Background on the Topic

The large margin concept has been hailed recently as the underlying principle that unites various learning-based techniques to classify data from examples. The basic tenet is that in addition to reducing the raw training error, e.g., number of misclassifications, in classification tasks, we must also maximize the margin, i.e., the confidence level of the classification with respect to the optimal discriminant hyperplane. In the case where the data in the *input space* cannot be separated by a hyperplane, a nonlinear function, Φ, maps the data into some other dot product space, i.e., the *feature space*. Although the difficulty seems to be finding the the explicit mapping, $\Phi(\mathbf{x})$, the Mercer Condition suggests that the solution of the dot products can be evaluated by a simple kernel function instead. Although the work of Vapnik, Mangasarian, and others laid the theoretical foundation of the large margin classifiers (LMCs) in the 1960s, we are seeing a revival of interest in LMCs, kernel machines, and other related topics due in part to the availability of better theoretical analysis tools and also the need of advanced computational techniques for large scale classification problems.

Background on the Volume

The volume titled *Advances in Large Margin Classifiers* by Smola *et al.* is the result of a two-day workshop on the topic of LMCs at the annual Neural Information Processing Systems (NIPS) conference in 1998. This volume contains mainly a number of manuscripts from the workshop and a few additional chapters from renowned leaders on the subject. In fact, the author list reads more like a Who's Who of luminaries in the subject field. Many are from the academia community with varying backgrounds in mathematics, statistics, computer science, electrical engineering, and physics, while the others are from commercial and governmental research laboratories. With this mixture of expertise, the volume offers a good balance between theoretical issues and algorithmic concerns that target both researchers as well as practitioners alike.

Overview of the Volume

The volume contains twenty-one chapters, which is further divided into six major sections: the preface, support vector machines (SVMs), kernel machines, boosting, leave-one-out methods, and beyond the margin.

The preface section introduces the fundamental concepts of LMCs and summarizes the volume with a detailed roadmap in Chapters 1 and 2, respectively. The SVM section, Chapters 3–7, is a more detailed treatment of various aspects of the SVMs. The next section on kernel machines, Chapters 8–10, takes on a more macroscopic view by examining the generalized SVM and how SVMs are related to other algorithmic and regularization networks. The volume then shifts back to examine Boosting more closely in Chapter 11 to 13. The leave-one-out methods section, Chapter 14–17, looks at the leave-one-out methods and other variations used in SVMs. The volume then concludes with the section on beyond the margin, Chapter 18 to 21, by presenting several novel views on the subject. In the following, we will organize the review according to the major sections with a snippet description on each chapter.

The preface section gives a good overview of the LMCs. The introduction, Chapter 1, is well illustrated with several pieces of pseudo code for readers to gain a better understanding of the topic. Key concepts (index terms) are also clearly indicated in the margin of the book for easy reference. Furthermore, the authors give web links to implementation and other resources on the subject matter. A list of mathematical symbols used in the book is also included to bring consistency to subsequent chapters. Chapter 2, the roadmap, offers a clear layout and overview of the volume from the editors' point of view, highlighting the contribution of each chapter.

On SVMs

The five chapters in the SVM section focus primarily on the extension of SVMs. Watkins examines how to utilize dynamic alignment algorithm for sequences as kernel functions in Chapter 3. In Chapter 4, Oliver *et al.* provide a regularization-theoretic analysis of the Fisher kernel based on generative models to demonstrate that the Fisher kernel equates to a L_2 norm regularization. Since standard SVMs do not provide posterior probability, Platt, in Chapter 5, extends the SVMs by mapping the SVM outputs into probabilities. In Chapter 6, Kowalczyk introduces a modified local learning rule derived from the classical perceptron that will converge to the optimal hyperplane resulting in the maximal separation margin. In Chapter 7, Herbrich *et al.*present a large margin algorithm for ordinal regression where preference relations between pairs of objects are used.

On Kernel Machines

In the kernel machines section, the three chapters examine the SVMs in a broader context. For instance, in Chapter 8, Mangasarian shows how the generalized SVM algorithm can use arbitrary kernel functions that do not satisfy the Mercer's condition. Guyon and Stork examine various relationships between different linear discriminant techniques in Chapter 9. In Chapter 10, Evgeniou *et al.* propose a unified framework for both regularization networks and SVMs.

On Boosting

Boosting and related ensemble learning have demonstrated their ability to improve the performance of classifiers dramatically. In Chapter 11, Rätsch *et al.* discuss how to use a single free parameter to determine the fraction of margin errors. Consequently, users with a rough estimation of the expected error may specify this parameter to reduce the training time. In Chapter 12, Mason *et al.* present a generalization of boosting algorithms by finding linear and convex combinations of functions that minimize arbitrary cost functionals. In Chapter 13, Karakoulas and Shawe-Taylor propose a new strategy for boosting learners for regression.

On leave-One-Out Methods

In the leave-one-out methods section, four chapters related to leave-one-out and cross validation methods for SVM are presented. In Chapter 14, Vapnik and Chapelle focus on using the leave-one-out estimator to derive bounds on the expectation of error. In Chapter 15,Weston and Herbrich modify the original SV algorithm by maximizing the individual margin adaptively and obtain competitive results when compared to SVMs and LP-SVMs (linear programming). Wahba *et al.* propose the generalized approximate cross validation (GACV), which can be used as an estimate to minimize the generalized

I. King is with the Chinese University of Hong Kong, Hong Kong.
R. Rojas is with Freie Universität Berlin, Berlin, Germany.
W. Dubitzky is with the School of Biomedical Sciences, University of Ulster, Coleraine, U.K.
Digital Object Identifier 10.1109/TNN.2003.814962

comparative Kullback-Leibler distance (GCKL), in Chapter 16. An overview of the relationship between Gaussian processes (GPs) and SVMs is presented in Chapter 17. Moreover, Opper and Winther use the "naive mean field" approximation to provide the posterior mean for GP classifiers and a linear response method to improve the computation of the generalization error of the leave-one-out estimator.

On Beyond the Margin

The purpose of the Beyond the Margin section is to introduce novel and outside-of-the-box concepts. Chapter 18 begins with a ray-tracing algorithm for estimating the Bayes classifier by Ruján and Marchand. In Chapter 19, Shawe-Taylor and Cristianini show how margin distribution and soft margin can be used to bound the generalization errors. Using techniques from statistical mechanics, Dietrich *et al.* examine the generalization performance of SVMs in large dataspaces in Chapter 20. In the final chapter of this volume, Smola *et al.* provide functional analytic bounds on the entropy numbers for convex combinations and multilayer perceptrons.

Summary

This volume has achieved what the editors set out to do, namely present an initial examination of LMCs. The strength of the book is in the breadth that it tries to cover in terms of the various aspects of LMCs and related topics. The editors have made great efforts in bringing a well-balanced and well-thought-out volume that offers much more than a collection of manuscripts from the workshop. This volume extends a typical workshop proceedings in several meaningful ways:

1) Key terms (concepts) are clearly identified in the margin of the book for easy reference.
2) Consistent mathematical notation is used across chapters which gives the whole volume a degree of consistency and coherence that no doubt have made the subject easier to digest.
3) There is a single unified reference section at the end of the volume listing almost 300 entries on work relating to the topic. This is invaluable for anyone who is working on the subject.
4) The editors have wisely made inter-related references among chapters to complement the understanding of the topic.

All these contribute significantly to the understanding of this piece of work.

One thing that leaves me wanting is the Index section. Although key concepts are being used as index terms, not all chapters and not all key concepts are being indexed at the end. Nonetheless, this collection is an authoritative volume that renders an excellent overview on the exciting and emerging subject. In short, *Advances in Large Margin Classifiers* provides a great resource for all researchers and practitioners a first-hand look at the state-of-the-art theoretical analyzes and algorithm designs on LMCs.

Pattern Classification, Second Edition—R. O. Duda, P. E. Hart, D. G. Stork (New York: Wiley, 2001, 654 pp.) *Reviewed by Raúl Rojas*

Duda and Hart's original book of 1973, *Pattern Classification and Scene Analysis*, is one of the most cited in the field of pattern recognition but had not been available in print for several years. The second edition of this evergreen updates the book in many respects (introducing at the same time a third author) and makes it widely available for new generations of researchers and students. In the past, the style and approach of "Duda-Hart" was adopted by other books but the original of 1973 remained unsurpassed—until now.

Pattern Classification consists of ten chapters and an appendix of mathematical foundations. Each chapter intermixes theory with practical examples and pseudo-code for important algorithms. Problem sets, computer exercises and chapter summaries complement each section. All illustrations for the book are available through the Internet in PDF format. Instructors can download them for use in class. The book can be used for one or two semester courses, depending on the material selected.

After a short introduction the book presents in Chapters 2 and 3 the fundamentals of Bayesian decision theory and Bayesian parameter estimation. The basics of classification theory are discussed: discriminant functions, decision surfaces, error bounds, probability densities. Special more advanced sections have been added in every chapter of the second edition (marked with a star). For example, Bayesian belief networks are introduced right at the end of Chapter 2. In Chapter 3, after discussing the computation of conditional densities, the reader learns to find more complex densities applying the Expectation Maximization (EM) algorithm. A section on hidden Markov models closes this chapter.

Chapter 4 deals with nonparametric techniques and the computation of densities using the experimental data. Parzen windows and k-nearest neighbors are the classical techniques, supplemented now in this second edition with a section on fuzzy classification.

Chapters 5 and 6 are devoted to the classification of patterns using linear techniques and multilayer neural networks, respectively. Chapter 5 goes much deeper than in other books. In many cases, a problem can be solved using linear classifiers before trying to apply more complex nonlinear ones. The advantage of linear classifiers is that they have a well-understood theory and are usually fast. But the authors also try to go beyond the classical methods and introducing modern material at the right place. This is the case of the section on SVMs, which are linear classifiers in a transformed space and nonlinear classifiers in the original space where the data comes from. Although the section is not very long, it provides the necessary insight into the general approach and sufficient pointers to literature to enable readers willing to learn more about the subject to go to the pertinent books. The chapter on neural networks, the nonlinear classifiers, is comprehensive and contains many subsections to help the reader speed-up convergence of learning algorithms such as backpropagation. Second order techniques and several classes of neural networks (recurrent, convolutional, etc.) are also covered.

Chapter 7 deals with stochastic methods for pattern classification (simulated annealing and related approaches) including genetic algorithms and genetic programming. I have a feeling that since genetic methods are quite generic and can be used for any kind of optimization task they could have been spared to make the book more compact and avoid "information overload".

Chapter 8 abandons the realm of metric classifiers and deals with machine learning in nonmetric spaces and well-known classification techniques such as ID3 and its cousin C4.5. These kind of classification problems have become more prominent now that large databanks of text or biochemical information exist. String comparison, for example, is used in the Internet and for the genome project. Good classification methods for this kind of problems are still being studied.

Chapter 9 deals with techniques which are algorithm-independent, such as resampling and the bootstrap method. The reader learns to apply the minimum description length (MDL) principle and just what it means. The chapter could have been placed earlier in the book and the instructor of a course could even spread this material over all chapters. Bootstrap techniques are useful for any kind of classifiers and MDL is useful for the training of neural networks.

Chapter 10, on clustering, closes the book. Statisticians might not be very happy about placing the unsupervised learning techniques at the end of the book, but at least they are there. Several different methods for clustering, including hierarchical clustering and the self-organizing feature maps are dealt with in this part of the book.

The new Duda–Hart–Stork is very comprehensive. The authors succeed in making the book a "one stop" warehouse of information on

pattern classification. The book is well written, the graphics and diagrams are superb and help the reader to get geometrical insight.

Who is likely to benefit from reading *Pattern Classification*? Certainly all computer scientists and engineers interested in artificial intelligence and also those who have to deal with pattern recognition problems in the fields of handwriting or speech recognition, industrial automation, computer vision, robotics, etc. The range of problems in which pattern recognition techniques are applied has grown considerably since 1973 and the second edition of *Pattern Classification* will probably have as strong an impact in the pattern recognition community as the first edition had.

Advances in the Evolutionary Synthesis of Intelligent Agents—M. J. Patel, V. Honavar, and K. Balakrishnan (The MIT Press, 2001, 480 pp, hardcover, ISBN 0-262-16 202-6) *Reviewed by Werner Dubitzky*

This text presents a most comprehensive overview of state-of-the-art for synthesizing neural and evolutionary computing models. A running theme underlying most of the contributions are the working assumptions of artificial intelligence and artificial life. Informally, these assumptions hold the view that thought and life processes can be understood and modeled by computation or information processing. In addition to this general background, the contributions reflect the realization that systems whose information processing structures are programmed to their last detail are extremely difficult to design for all but the simplest of applications. Within the framework of these general assumptions, the first chapter sets out to introduce the powerful agent synthesis paradigm of evolutionary and neural learning. From this introduction Balakrishnan and Honavar continue to develop a set of dimensions or properties critical for studying and evaluating this new paradigm. The outlined critera turn out to be an excellent backdrop and useful guideline for reading and studying the remainder of the book.

Chapters 2–5 deal with the evolutionary synthesis of complex agent programs. The tasks addressed include walking gait modeling, evolution of agent communication, unsupervised classification, and emergence of box-pushing behaviors. Gruau and Quatramaran present an interesting interactive or semi-unsupervised approach that impressively demonstrates how human intervention can be efficiently exploited in identifying and promoting the emergence of specific regularities used by a neurocontroller. Arming agents with a final state machine, McLennan demonstrates that the evolved agent populations appear to communicate meaningfully using in some cases rudimentary syntax. Honavar and Balakrishnan address the evolution of effective box-pushing behavior. They use multi-level, variable-length genetic representations to evolve feed-forward, recurrent neurocontrollers. Besides highlighting the role of memory, they also show how the learned structures can be meaningfully analyzed and understood so as to gain new insights for future design of such systems. To this end, their study is highly educational to researchers working in the field.

Chapters 6–8 explore models that are inspired by mechanisms found in the biological development of organizms. Testing their models on a variety of tasks, the authors demonstrate the benefits of the developmental approach both in terms of effectiveness and efficiency of the resulting systems. Boers and Sprinkhuizen–Kuyper, for example, demonstrate the scalability in terms of coding large neural-network structures on the basis of a parallel string rewriting mechanism. Michael's approach to developing large recurrent neural networks based on a neurogenesis model demonstrates that this approach has great potential for designing large and highly modular networks. He also lives up to the promise of artificial life studies to actually use the insights gained from in silico experiments and apply them to real-world problems. In this case, he transfers the agent behavior evolved in simulation to real robots.

The contributions in Chapter 9–11 investigate the power of local search algorithms in conjunction with evolutionary search to further shape the solutions discovered by evolution. Carse *et al.*, for example, explore a hybrid approach to discover optimal network structure and an learning algorithm for RBF neural networks. The authors present two novel mechanisms to make search more efficient and to home in on promising network structures early in the learning process: 1) a novel crossover operator for identifying similar units and 2) a method for the scaling of training epochs. They demonstrate that for a number of problems this approach is able to produce relatively small yet accurate RBF structures. Within the context of chase-escape behaviors, Floreano *et al.* study robots, controlled by recurrent networks, that operate in a coevolutionary setting. Their investigations are important as they are inspired by the thesis that unpredictable and dynamic environments require evolution of adaptive behavior, a critical tenet underlying agent systems research. To test evolving agents in dynamic fitness landscapes, the authors explore mechanisms such as preferential genotype selection, evolution based on constant-diversity populations, and ontogenetic adaptation. They present and discuss interesting emergent behaviors that they observed in the experiments. Clearly, their investigations are an important stepping stone for more research of this kind.

The final chapters of the book (Chapters 12–15) are arguably the most challenging ones, as they address some of the fundamental issues of neuro-evolutionary algorithms. Among them are the phenomenon of premature convergence and the lack of concrete theoretical underpinnings. An interesting approach presented by Menczer *et al.* puts forth the idea of local selection and constraints on population diversity. Using a direct coding scheme on feedforward nets, the authors demonstrate the power of their approach on a number of application domains. Another inspirational new way of thinking about neurons is presented by Moriarty and Miikkulainen. In their scheme, each individual in the population is a neuron, rather than a neural network. The authors demonstrate the benefit of their approach with respect to premature convergence. Further contributions in this part of the book discuss evolutionary programming approach to evolving neural networks and network ensembles, and a theoretical study analyzing genetic algorithms from the perspective of population genetics and statistics. The results of the former study suggest that network ensembles are often outperforming the best-performing single network. And the latter study shows, among other things, that genetic search can be closely approximated by a new algorithm called univariate marginal distribution algorithm.

The book is clearly written and can be recommended to anyone interested in the synthesis of evolutionary and neural computation, and agent-based systems. Favoring comprehensive chapters over more but less detailed chapters, the volume does provide a precious service to those sincerely interested in the technical discussion. However, as the chapters proceed in an easy to follow top-down fashion, those interested in an overview of the various techniques are served equally well. Throughout the book, a well-balanced treatment of theoretical and applied aspects of the discussed material is maintained. This makes the volume attractive to a wider range of readers. Most of the presented approaches are evaluated on a set of diverse problems. By doing so, the developers of the underlying methods provide the reader with a valuable "handle" to an in-depth comprehension of the discussed schemes. In many respects, the work on an agent synthesis of evolutionary and neural learning is still in its infancy. This book presents one of the rare excellent texts on the subject matter, it constitutes a stepping stone to future development in this field.